DOCUMENT RESUME

ED 068 501 TM 001 838

AUTHOR Young, Jon I.
TITLE Model for Competency-Based Evaluation.
PUB DATE [72]
NOTE 16p.

EDRS PRICE MF-$0.65 HC-$3.29
DESCRIPTORS *Evaluation Methods; *Measurement Instruments; *Models; Performance Tests; Research Methodology; Technical Reports; *Test Reliability; *Test Validity

ABSTRACT

Some theoretical concerns for competency-based evaluation instruments are discussed, and means of examining these instruments for validity and reliability are presented. The areas of concern include descriptions of the behavior, the level of response, and the nature of the evaluation. Two different types of instruments are examined to determine the validity and reliability of each in directly and indirectly evaluating a process. The first instrument indirectly evaluates an ability to classify four types of questions: memory, convergent, divergent, and evaluative. The data were collected by an objective test with a standard criterion. The second instrument, a rating form, measures microteaching (a process) directly without a standard criterion. The data indicate that these instruments have potential for giving accurate evaluations. (Author/LH)

MODEL FOR COMPETENCY-BASED EVALUATION

Jon I. Young

Brigham Young University

ABSTRACT

The paper discusses some theoretical concerns for competency-based evaluation instruments and presents how these instruments may be examined for validity and reliability. In addition, two different types of instruments were examined to determine the validity and realiability of each.

MODEL OF COMPETENCY-BASED EVALUATION

Jon I. Young

Brigham Young University

In conventional testing situations, students respond to questions which are designed to measure their academic knowledge about some subject. In competency-based evaluation, each student is expected to demonstrate his ability to deal with situations that are as close as possible to real-life situations. This paper will identify and discuss some of the critical concerns with competency-based evaluation.

Description of the Behavior

Having once identified the expected outcome to be evaluated, it is important to specify whether the behavior is a process or a product. Behaviors identified as processes must be evaluted during the performance of the behavior. However, if the process can be logically or empirically tied to some product, the process may be evaluated indirectly by evaluating the product.

If the behavior requires that the student will produce some product, then the behavior may be evaluated any time after it occurs. An evaluation that occurs simultaneously with the behavior, or examines a product, is called direct evaluation. An indirect evaluation is employed when direct observation is not possible or practical. This method evaluates a dependent product and inferences to the actual behavior. Figure 1 illustrates the two behaviors and interaction between them. In each of the four categories an example is given which helps to illus-

trate the type of interaction envisioned.

What is to be evaluated? / Mode of evaluation?	Doing (Process)	Done (Product)
Direct	A	B
Indirect	C	D

Figure 1. Evaluation Model in Competency-based Education

(A) Microteaching
(B) Writing Program
(C) Report on a case study
(D) Representation of a product

Description of the Response

Before evaluation instruments can be appropriately designed, the nature of the response must also be identified. Although there are numerous ways in which a student may show his competencies, the two most common are verbal and written. The evaluation must then agree with the mode of responding. For instance, if the behavior is for the student to write a peom, the evaluation should not examine his ability to recite a poem to the class.

Level of Responding

It is especially important that the expected behavior be identified in

terms of the appropriate cognitive level (i.e., memory, conceptualization, principle using, problem solving; Gagne 1965). A valid measuring instrument will not evaluate the response at a level different from the level at which it occurs. An accurate assessment of the student's ability to solve a social problem should not require that he reproduce or recognize elements of it, although these abilities may be prerequisites to solving the problem.

Special consideration is given to the psychomotor domain and the affective area. Whereas the evaluation of cognitively oriented behaviors may occur with appropriate sampling of possible responses, psychomotor behaviors are often evaluated in their entirety.

The affective domain evaluation relies almost exclusively on indirect evaluation. Affective behaviors are often considered process behaviors because of their inaccessibility to direct measures. However, affective behaviors can be specified and evaluated almost as objectively as cognitive behaviors (Lee, 1972).

Levels of Evaluation Criterion

Turner (1971) specifies six levels of sophistication that must be considered in selecting or designing evaluation instruments. These levels begin with the easiest type of assessment, the evaluation of knowledge, and progress to a level that involves both direct and indirect evaluation.

At Level 6 the student need not engage in a process or product but only show "understanding" for some concept or principle which is probably essential, or enabling, to a higher level of responding. "Understanding," in this context,

may be defined in numerous ways (Bloom, 1956).

Level 5 requires the student to demonstrate that he possesses a particular skill, usually by reacting to some prepared material.

Level 4 is evidence that the student can perform his skills or apply his knowledge under very restricted but real conditions. He must demonstrate his ability with real problems.

Level 3 uses more liberal and expanded conditions than Level 4. Here the student must show his ability in the actual situations he will face. However, he is not expected to rely totally on his own resources, but may perform in consultation with a supervisor, and the behavior usually occurs in a pre-organized situation.

Level 2 begins the final step in evaluating competencies and specifies that the student will demonstrate his ability in a real situation over a set period of time. During this period, the student will be evaluated directly on process behaviors and directly on the products these processes produce. Thus the student evaluation is two-fold.

Level 1 is exactly the same as level 2, but over a longer period of time. During this time the student's ability is confirmed and the student's subsequent performance may be accurately predicted.

Turner's taxonomy can be used to further specify the nature of the evaluation. Levels 5 and 6 evluate student processes by indirectly evaluating a process or product, (i.e., knowledge), generally by a test. The real usefulness of these two levels is to permit some inferences about the adequacy of the learning sequences.

Levels 3 and 4 rely on direct evaluation of process behaviors. Levels 1 and 2 use both direct evaluation of process behaviors and direct evaluation of products of those processes. Obviously, the levels are specifically designed to evaluate summatively the usefulness of instructional activities in respect to particular behaviors or skills. When used as a complete evaluation system, the levels can collect normative data which may be useful in counseling students in view of their respective performances.

Having described the level of sophistication of the evaluation and what is expected to be evaluated, the next considerations are to be directed toward the internal nature of the evaluation instrument.

Evaluation Description

One of the major concerns must be directed toward the type of response that is expected. There are normally two types of evaluations that are used in competency-based measurement. First is the evaluation for which a key of correct responses is available. This type of evaluation is most commonly used to measure a process indirectly (like a test) and occurs in levels 5 and 6.

When this type of measurement is used in the psychomotor or affective domain (and possibly in the cognitive domain), the standard evaluation is an "all or none" situation. If the behavior occurs it has met the standard. Evaluations of this kind are only possible when the student reacts to prepared materials, or the total number of specified responses is small enough to be specified.

The second type of evaluation response is a rating scale from an objective evaluator. These ratings normally occur when evaluating a process, but

may also be used to evaluate a product that has no standard of excellence. These evaluations are useful because of the inherent variability and sensitiveness to individual differences. Any number of aspects of the process or product can be examined and at any degree of sophistication. These evaluations are mostly used below level 5.

Evaluation must also consider the level of analytic sophistication possible. This concern derives mostly from the need to truly sample the behavior. Often the task may be so complicated that only a small sampling can be made, thus limiting inferences about student ability. If the behavior is truly comprehensive, the subordinate behaviors must also be evaluated, often done most practically at another level or in another domain. This concern is especially critical when dealing with evaluations of standard responding.

Validity

It is the right and obligation of the evaluator to decide which of the validations are pertinent to the instrument being used. Face validity is the most common form of specifying the adequacy of the least effort the designer can undertake. Face validity is most concerned that the instrument be in agreement with the mode of responding (written or verbal), whether it measures a product or a process and the level of responding (memory, conceptualization, etc.).

Content validity may be estimated by expert ratings of each test item. If the behavior can only occur in one manner, content validity is safely assumed. However, if the behavior can be measured by more than one item, it must be

shown that the item used is a random selection from all possible items that could be used.

Construct validity is estimated by giving the evaluation to a group of persons possessing the knowledge or ability and to a group not possessing the attribute of interest. If the "expert" group performs significantly better, the instrument may be considered contruct valid.

Predictive validity is best determined between levels of the evaluations (Turner's hierarchy), or at the same level but in different time periods.

Reliability

Estimates of reliability can be obtained by several standard procedures. The Spearman-Brown technique is a means of estimating the reliability of a short test assuming more items were potentially available. However, if the evaluation is one that is an "all or none" instrument, the Spearman-Brown technique is not appropriate, since the total universe is already specified and sampled.

It must be recognized that using the Spearman-Brown procedure is at best an artificial technique and should be employed sparingly to estimate the reliability of the test being used. In short, this technique estimates reliability under perfect or near perfect conditions and not the practical constraints of the actual test.

Using a split-half procedure permits an estimate of the reliability by comparing subsections of the test. Both sections are compared as if they were separate or parallel forms of the same test, and a simple correlation is computed. Using this procedure assumes that at least two items in the test will measure the same behavior or concept.

The final measure of reliability to be considered is that of a test-retest procedure. This measures the stability of the evaluation and is usually accomplished by computing the correlation between the first and second administration of the same test. Generally a time period of about two weeks separates the two administrations.

If the evaluation does not have a standard of measure, the test-retest coefficient occurs between two or more ratings by the same evaluator. The internal consistency of the measure may be computed with a correlation between the ratings of two or more evaluators taken at the same time.

Conclusion

Competency-based evaluation is a necessary concern in today's educational processes. If adequate care is taken in describing the behavior, the level of responding and the nature of the evaluation, then appropriate instruments can be designed.

Once the responses and evaluations are specified any standard procedure for estimating validity and reliability may be implemented using the appropriate data.

DATA COLLECTION

Two types of evaluation instruments were used in an individualized teacher training program (ISTEP) at Brigham Young University and were examined to estimate their reliability and validity. The instruments were used to directly and indirectly evaluate a process. The first instrument will indirectly evaluate an ability to classify four types of questions; memory, convergent, divergent, and evaluative. The data was collected by an objective test with a standard criterion.

The second instrument measured microteaching (a process) directly without a standard criterion. This instrument was a rating form.

Study # 1 Indirect evaluation of a process with a standard criterion

Procedure

A twenty-four item (six items for each categroy of question) test was constructed and given to ten experts and ten naive subjects. After studying a series of instructional materials, the test was given to a large group of subjects and was repeated one week later.

Results

The results showed a significant difference between the expert and naive subjects ($t=8.575$, $p < .001$). A correlation between the experts on each of the twenty-four items ranged from 1.00 to .00.

A correlation between the two testing sessions showed a significant correlation ($r=.6843$, $p < .001$) between those who passed the test

both times and those subjects who failed the test both times. There was also no significant difference between the mean scores of persons taking the two tests ($t=.552$, $p>.05$). However, there was an insignificant rank order correlation because the exact order of each person's score was not the same.

A factor analysis showed eight factors were involved instead of the predicted four factors. The items associated with each of the eight factors were correlated at least .35.

Finally, the test was divided into halves so that each category was measured by three items for a total of twelve items. The rank order correlation was not significant but the mean scores for the two halves were almost identical (10.40 and 10.50).

Validity

Content validity was perfect for twelve of the items and random for the other twelve. Therefore, the test should be modified to use only the twelve valid items, or change the others until they become acceptable.

Construct validity was shown by the significant difference between the expert and naive scores. Additional construct validity was measured by a factor analysis which showed eight factors. Those items which are associated with each factor would have to be evaluated and their similar characteristics specified to determine exactly what the factor was measuring.

Reliability

A split half reliability was estimated by dividing the test in half. Although the correlation between scores was not significant, the mean scores for the two halves were almost identical.

Finally, a correlation was computed between the two tests, showing an insignificant rank order relationship but a significant relationship between those subjects passing both tests and those failing both tests.

Summary

Although the estimates of validity for this test were fairly positive, the reliability was accurate only between those passing the test and those failing the test.

Predictive validity and concurrent validity were not concerns with this instrument, although there may have been some prediction for the application of these questions in real situations.

The factor analysis indicated some information that can only be useful when the items of each factor are analyzed to determine what are the characteristics of each factor.

Study #2 Direct evaluation of a process without a standard criterion

Procedure

Three experienced evaluators were asked to rate the performance of each of three pre-service teachers. This rating occurred twice with a week's separation. The ratings between the pre-service teachers for each

evaluator were correlated, as well as the ratings between the evaluators for each pre-service teacher.

Results

The inter-rater correlation was significant in all three evaluations. The first evaluation was positively significant ($r=.5705$, $p<.05$) for raters #1 and #2. The correlation on the second evaluation was significant ($r=.9533$, $r=.9955$, $p<.001$) between all three raters. The correlation on the third evaluation was significant between evaluators #1 and #2, but in a negative direction.

The intra-rater correlations were .49, .59, and .69, not including those items which were perfectly correlated. Two of these (.59 and .69) are significant at the .05 level.

Two factor analyses were computed between (1) the raters and (2) the students. In both cases two factors were found. The first factor accounted for a substantial amount of the variance (.854 for the raters and .740 for the students), indicating that the items contributing to this factor were strong evaluation tools. The items correlated with these factors from .701 to .999.

Validity

Content validity was measured by having a panel of experienced evaluators (5 persons) rate the individual items in terms of their appropriateness. This task basically involved examining the items to see if

they communicated clearly, because they were "all or none" measures. The panel's scores correlated highly (r=.9382).

Construct validity was measured by factor analyzing the instrument. The results showed two strong factors with the items correlating fairly high (.70) to very high (.99). These two factors appeared consistent, and items associated with each were constant.

Reliability

The inter-rater reliability was good ($p < .05$), showing that different experienced raters evaluated each student's microteaching consistently.

The intra-rater realiabilities were also significant ($p < .05$), indicating that the evaluations are potentially stable.

Summary

The data from this study indicates that the instrument is appropriate to use in evaluating microteaching performance. Some of the items gave little information and would need to be revised. The items in each of the two factors should be examined to determine, if possible, what is being measured. However, the instrument is consistent in its evaluations.

Predictive validity and concurrent validity were not examined. However, this could be done using another evaluation form (concurrent) and comparing this evaluation with later performance using the same form (predictive).

DISCUSSION

Although the instruments examined were not exact, these procedures can be used to gain accurate information about evaluation tools prior to investing large amounts of time, effort, and expense.

The data indicate that these instruments have potential for giving accurate evaluations. Now the individual items of each can be examined to determine whether they are adding or detracting from the appropriateness of the instruments.

REFERENCES

Bloom, B. S. Taxonomy of Educational Objectives: The Classification of Educational Goals, Handbook I: Cognitive Domain. David McKay Company, Inc., New York, 1956.

Gagne, R. M. The Conditions of Learning. Holt, Rinehart, & Winston, Inc., New York, 1965.

Lee, B. N. The development and validation of two forms of a branching program for writing affective objectives (Unpublished Thesis). Brigham Young University, 1972.

Turner, R.L. Programmatic Themes and Mechanisms. The Power of Competency-based Teacher Education. Rosner, B. (Editor). Project No. 1-0475, U.S. Office of Education, 1971.